The Connecticut Academic Performance Test: Technical Report

Prepared by
Irene Hendrawan & Arianto Wibowo

January 2011

Table of Contents

List of Charts

List of Tables

Part 1: Introduction

1.1. General description of CAPT

The Connecticut Academic Performance Test (CAPT) was designed to measure student performance in high school. Students are tested in the areas of Mathematics, Science, Reading, and Writing.

The CAPT has measured achievement of Connecticut students since 1994, when it was first administered. The second generation of CAPT was introduced in 2001. The content structure of the first generation CAPT was used as the baseline in developing the second generation. For the most part, the educational outcomes tested in the first generation were carried over to the second generation. Changes were made in light of new trends in instruction, educational assessment, and the lessons learned over the years of the first generation. The third generation of CAPT was introduced in the spring of 2007. The spring 2010 administration was the fourth operational (OP) administration of CAPT3.

1.2. 2010 CAPT Test Design

The spring 2010 administration comprises the following content areas:

1. Mathematics
   Mathematics (MA) has thirty-two operational items -- twenty-four grid-in (GR) response items and eight open-ended (OE) items scored on 0-3 scale.
2. Science
   Science (SC) has sixty-five OP items -- sixty multiple choice (MC) items and five OE items scored on 0-3 scale.
3. Reading
   Reading (RD) consists of two subtests:
   - Reading for Information
     Reading for Information (RI) has eighteen OP items -- twelve MC items and six OE items scored on 0-2 scale.
   - Response to Literature
     Response to Literature (RL) consists of an extended response (EX) item with a 2-12 score scale (sum of two rater scores on a 1-6 scale).
4. Writing
   Writing (WR) consists of three subtests:
   - Editing & Revising
     Editing & Revising (ER) has eighteen MC items.
   - Interdisciplinary Writing 1 & Interdisciplinary Writing 2
     Interdisciplinary Writing 1 (IW1) & Interdisciplinary Writing 2 (IW2) have EX items with a 2-12 score scale (sum of two rater scores on a 1-6 scale).

Table 1: 2010 CAPT Operational Test Design

Content Area	Subject	Number of Items				Total Items	Raw Score
		MC	GR	OE	EX		
Mathematics	Mathematics		24	8		32	0 - 48
Science	Science	60		5		65	0 - 75
Reading	Reading for Information	12		6		18	0 - 24
	Response to Literature				1	1	2 - 12
Writing	Editing & Revising	18				18	0 - 18
	Interdisciplinary Writing 1				1	1	2 - 12
	Interdisciplinary Writing 2				1	1	2 - 12

1.3. 2010 CAPT Test Forms

In the 2010 administration, two main forms were available for administration: Form HS17, which is the live form taken by most of the students, and Form HS0, which was available for breach situations. Moreover, Form HS0 will be used as a breach form in subsequent years of the third generation. Although the two forms were pre-equated during test assembly, there was still a need to carry out a post equating procedure after the test administration in order to ensure the comparability of the two forms.

CSDE's rationale for stratifying the test forms based on scale scores from the previous year was that this procedure would more likely yield groups of test takers who were representative with respect to the distribution of skills and achievement across the entire state. In other words, instead of sampling based on conventional demographic variables to achieve representation of test-taker characteristics, CSDE chose to sample on test-taker achievement. MI selects a stratified sample of schools, based on the scale score distribution to which each belongs.

Any student who breaches a test session or subtest (HS17 or HS0) was given the corresponding test session or subtest (HS17 or HS0).

Part 2: Test Development

The process by which each form of the CAPT is developed is extensive, spanning a five- or six-year period and many stages. The development process is led and overseen by staff members in the Bureau of Student Assessment at the Connecticut State Department of Education (CSDE), but it also involves many other people who represent a wide variety of perspectives and areas of expertise. CSDE curriculum specialists and content experts play a critical role and work closely with the assessment staff throughout the process. In addition, a major testing company and other organizations and individuals with experience in educational assessment are involved at appropriate points in the development process.

Advisory committees of Connecticut educators are particularly important throughout the development of the CAPT. Content Advisory and Fairness Committees review each item to ensure the match between the content objectives and the items, and to ensure meaningful interpretability of test results. The Content Advisory Committees included content experts, regular and special education teachers, Connecticut State Department of Education curriculum, and assessment content specialists. A separate advisory committee is established for each part of the CAPT: Mathematics, Science, Reading, and Writing. These advisory committee members are selected on the basis of their knowledge in educational content and processes. In addition, the Fairness Committee is responsible for determining whether items are appropriate and fair to all examinees. Educators are carefully selected for the advisory committees to be representative of school districts throughout Connecticut.

The test development process for CAPT3 began with content specialists and testing experts writing test specifications with the help of the CAPT content advisory committees. The starting point for this process was looking at the specifications and structure of the first generation CAPT, and examining what has been working and what needed improvement. The new curriculum frameworks adopted by the State of Connecticut were also used as a guide. Test items for the CAPT3 were carefully developed in accordance with the established test specifications and test blueprint. These items were carefully matched to the content standards in the Connecticut Curriculum Frameworks for Mathematics, Science, Reading, and Writing. Items that did not pass the scrutiny of either Content Advisory or Fairness Committees were eliminated from the pool of pilot items.

After committee reviews, field test forms were created and piloted on a representative sample, stratified by scale score distribution, consisting of approximately 2000 students per form. Pilot statistics such as the mean, point biserial, and Rasch difficulty were generated and reviewed by CSDE assessment content staff and psychometricians. In addition, for hand-scored constructed response items, the contractor staff provided qualitative summaries about whether students appeared to have sufficient contextual knowledge to be able to fully respond to the item. Flawed items were removed from the item pool, including those showing test item bias or inappropriate levels of difficulty. Based on the CAPT3 Blueprints, Mathematics, Science, Reading, and Writing test forms of equivalent difficulty were simultaneously constructed from the pool of items that met all the review criteria. Every effort was made to ensure that strand level difficulties were comparable and that the items reflected the appropriate range of content within the strands across the generation.

Part 3: Item Level Statistics

Table 2 and Appendix A present a summary and detailed result of item analysis (item quality) data for Mathematics, Science, Reading and Writing, respectively. The following information is presented in each item analysis:

Classical and IRT difficulties: Item difficulty is fundamentally a ratio of the proportion of examinees who answered the item correctly. Thus, an easy item has a high p-value and a difficult item has a low p-value. If an item has a very high p-value it may be so easy that it does not provide much information about what most examinees know or can do, while an item with a very low p-value may be so difficult that it is beyond the range of what most students know or can do. Therefore, items with very high or very low p-values may be rejected, unless content relevance overrides that concern.

The IRT difficulty described here is the Rasch IRT model's item difficulty parameter. This parameter influences the probability of correctly responding to the item as defined by the Rasch IRT model. For a given examinee's ability, the higher the IRT difficulty, the lower the probability of responding correctly. Thus, an easy item has a low Rasch difficulty and a difficult item has a high Rasch difficulty.

Item Discriminations: The point biserial correlation or item-total correlations measure the strength of the relationship between the particular item score and the total test score. Thus, item discrimination reflects how well a particular item differentiates between high and low total test performers. When the correlation is high, examinees that do well on the item also tend to do well on the entire test and correspondingly, examinees that do not do well on the item also tend not to do well on the total test.

Distractor Frequencies: The proportion of students who answered each option (A-D, 0-3, and 2-12) are presented for the multiple-choice items, open-ended and extended response, respectively.

Table 2: Summary of Item Analysis Form HS17

Subject	Rasch		P-value		Point Biserial	
	Mean	Std	Mean	Std	Mean	Std
Mathematics	0.0537	0.7163	0.70	0.38	0.54	0.11
Reading for Information	-0.0472	0.8468	0.72	0.19	0.40	0.12
Response to Literature	0.8968		6.70		0.62	
Editing and Revising	-0.1386	0.7398	0.75	0.11	0.33	0.06
Interdisciplinary Writing	1.2312	0.0100	7.67	0.01	0.72	0.00
Science	0.0387	0.5830	0.70	0.36	0.39	0.10

Part 4: Scaling and Equating

4.1 2010 CAPT Linking Items

The 2010 CAPT Mathematics, Science, Reading for Information, and Editing & Revising tests were equated with the 2009 CAPT (HS16) subtests by embedding linking items. Linking items were counted toward students' scores.

The Live form of the 2010 CAPT (HS17) included:

- Mathematics – twelve linking grid items were embedded.
- Science – fifteen linking MC items were embedded.
- Editing & Revising – one passage with six linking MC items were embedded.
- Reading for Information – one passage with four linking MC items and two linking OE items.

Table 3 indicates the linking items used as well as their positions on the 2010 and 2009 tests.

Table 3: 2010 Embedded Linking Items

Content Area	Form HS17 Item Position	Form HS16 Item Position	Rasch Form HS16
Mathematics	5	5	-1.3601
	7	6	0.2743
	8	7	0.0941
	13	14	-0.7499
	15	12	0.6805
	16	16	0.2533
	23	23	-0.2089
	25	25	-0.4666
	26	26	-0.0275
	28	29	-0.4829
	30	30	1.3458
	32	32	-1.3576
Science	5	5	-0.9562
	9	9	0.4411
	13	13	0.8981
	19	19	-0.1917
	21	21	-0.0810
	26	26	-0.6101
	34	34	-0.8413
	35	35	-0.0075
	36	36	-0.3601

Content Area	Form HS17 Item Position	Form HS16 Item Position	Rasch Form HS16
	37	37	-1.3230
	46	46	-0.0368
	47	47	-0.2142
	48	48	-1.0201
	52	52	-0.2748
	53	53	0.0360
Reading	7	7	0.9721
	8	8	-0.0386
	9	9	-0.3423
	10	10	-1.1422
	11	11	-0.3833
	12	12	1.3093
Writing	1	1	0.7217
	2	2	-0.7174
	3	3	-1.4069
	4	4	-1.3103
	5	5	0.5253
	6	6	0.2052

4.2. Calibration Process

The CAPT 2010 tests were scaled and equated using the Rasch model. The WINSTEPS software was used to estimate the latent trait difficulty of each item on the test. WINSTEPS, written by Linacre (Mesa Press, 2005) was used to complete Rasch analyses. WINSTEPS is a WINDOWS-based program that is widely used for similar high stakes tests. WINSTEPS (based on the Rasch model), allows for the estimation of item difficulty for multiple-choice, open-ended, and extended response items on a single scale. Using these item difficulties, the model is able to estimate the ability (theta) of each student corresponding to each student's raw score.

All scaling and equating analyses were undertaken by three independent groups: Measurement Incorporated (MI), the contractor, the Connecticut State Department of Education (CSDE), and H. Jane Rogers and H. Swaminathan from the University of Connecticut (UCONN). Results were compared and cross-checked to the fourth decimal point to ensure accuracy.

The purpose of equating was to place the difficulty estimates of the items on the same scale as HS16 (CAPT 2009 Live). The equating was accomplished in the following steps:

1. For every content area, calibrate all items in 2010 OP (see Charts 1-4 for sample calibration data matrix). This step is a free run calibration. For RL, IW1, and IW2, 2 is subtracted from each score so that scores are on a scale from 0 to 10.

Chart 1: Calibration Design for 2010 Mathematics

Form HS17	HS17_MA1	HS17_MA2

Note:
HS17_MA1 = Form HS17 Math Session 1
HS17_MA2 = Form HS17 Math Session 2

Chart 2: Calibration Design for 2010 Science

Form HS17	HS17_SC1	HS17_SC2

Note:
HS17_SC1 = Form HS17 Science Session 1
HS17_SC2 = Form HS17 Science Session 2

Chart 3: Calibration Design for 2010 Reading

Form HS17	HS17_RI	HS17_RL

Note:
HS17_RI = Form HS17 Reading for Information
HS17_RL = Form HS17 Response to Literature

Chart 4: Calibration Design for 2010 Writing

HS17	HS17_ER	HS17_IW1	HS17_IW2

Note:
HS17_ER = Form HS17 Editing & Revising
HS17_IW1 = Form HS17 Interdisciplinary Writing 1
HS17_IW2 = Form HS17 Interdisciplinary Writing 2

2. Select the items linking HS17 (2010 live test) and HS16 (2009 live test). Do anchor evaluation using .3 rule between the estimates of difficulties from Step 1 and HS16 values (see Table 3 for the Rasch values of linking items between Form 17 and Form 16). This is an iterative process in which each item, starting with the one with the greatest absolute value difference, is removed until all items fulfill the criterion for inclusion. Using the remaining items the difference between the scale means from HS16 and Step 1 yields the equating constant. Table 4 shows the equating constants for Form 17 and Form 16.

Table 4: 2010 CAPT Equating Constants

Content Area	Equating Constant
Mathematics	0.0537
Reading	0.0025
Science	0.0387
Writing	-0.0016

3. Using the item output files from step 1 and anchoring their b-values, perform another run for each combination of forms, i.e., employ only those items from a given form in order to obtain theta

values for each group of students administered a particular form. For Reading and Writing, the appropriate weights were included in the second calibration (see Table 5).

Table 5: Summary of Weighting for Reading and Writing

Content/Subject	Unweighted Scale	% of Total Scale	Score Weight	Compute Formula	Weighted Scale
Reading for Information	0 - 24	50%	1.0		0 - 24
Response to Literature	2 - 12	50%	2.4	(RL - 2)*2.4	0 - 24
Total Reading	2 - 36				0 - 48
Editing & Revising	0 - 18	30%	1.0		0 - 18
Interdisciplinary Writing 1	2 - 12	35%	2.1	(IW1 - 2)*2.1	0 - 21
Interdisciplinary Writing 2	2 - 12	35%	2.1	(IW2 - 2)*2.1	0 - 21
Total Writing	4 - 42				0 - 60

4. Compute scale score (SS) and scale score standard error (SSE) for each form:

$$SS = \left(\frac{T + EQ - T_{mean}}{T_{SD}} \right) * 45 + 250 \text{ and } SSE = \frac{T_{err}}{T_{SD}} * 45$$

where

T and T_{err} are the ability score and the standard error of the ability from the score file in Step 3 (for Reading and Writing) and Step 1 (for Mathematics and Science).

EQ is the difference between the mean of difficulty estimates of the linking items on HS16 and mean of difficulty estimates of the linking items on HS17, called the equating constant. This value was obtained in Step 2.

T_{mean} and T_{SD} are the scaling coefficients from base year of CAPT2 (see Table 6).

Table 6: Scaling Coefficients from Base Year (CAPT2)

Content Area	T_mean	T_SD
Mathematics	-0.2317	1.6051
Science	0.4077	0.9254
Reading	0.4843	1.2278
Writing	1.0931	1.1187

The minimum SS is set to 100 and the maximum SS is set to 400. Any SS less than 100 was reset to 100 and any SS greater than 400 was reset to 400.

Appendix B contains the results of raw scores, theta, and scale score for HS17. Please contact CSDE for other forms and combinations.

Part 5: Test Statistics

5.1. Reliability

Reliability is a statistical index of the consistency of test performance over repeated trials. The simplest model for conveying the concept of reliability is to describe the test re-test method. If a test is administered to a group of examinees and then re-administered to the same examinees a short time later, the correlation of the scores across both test administrations estimates the reliability of the test. To measure reliability using a single administration, the test items are split using various techniques into half-length tests and those scores are then correlated. Cronbach's alpha estimates the lower-bound estimate of an infinite combination of split-halves and therefore is regarded as a very conservative method for assessing test reliability.

Table 7 summarizes reliability estimates for CAPT Mathematics, Science, Reading, and Writing. The reliability coefficients are based on Cronbach's alpha measure of internal consistency. When evaluating these results it is important to remember that reliability is partially a function of test length and thus reliability is likely to be greater for tests that have more items. Table 8 presents the mean and standard deviation of students' scale scores.

Table 7: CAPT Cronbach's Alpha

Form	Mathematics	Reading	Writing	Science
HS17	0.936	0.829	0.802	0.929

Table 8: CAPT Scale Score Summary Statistics

Subject	Mean	Standard Deviation
Mathematics	250.16	42.22
Reading	246.36	59.25
Writing	263.43	62.89
Science	259.21	48.39

5.2. Classification Consistency and Accuracy

Classification consistency (see Table 9) and accuracy (see Table 10) were measured using the IRT-Class program developed by CASMA (Center for Advanced Studies in Measurement and Assessment) at the University of Iowa. The decision consistency and accuracy was assessed based on the given ability distribution and the difficulty of the items (IRT parameters).

Table 9: Classification Consistency

Content Area	Overall Classification Consistency	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.781	0.940	0.947	0.946	0.944
Reading	0.824	0.901	0.939	0.966	0.966
Science	0.785	0.958	0.954	0.938	0.927
Writing	0.924	0.961	0.967	0.964	0.967

Table 10: Classification Accuracy

Content Area	Overall Classification Accuracy	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.841	0.958	0.962	0.962	0.959
Reading	0.855	0.925	0.955	0.972	0.977
Science	0.844	0.971	0.967	0.956	0.949
Writing	0.915	0.956	0.979	0.973	0.977

The results of the program show that for the most part, classifications are highly consistent (see Table 9). The consistency ratings at each cut score are generally in the upper 90s. This tends to tail off at the highest cut score (i.e., the upper end of the distributions). The cumulative effect of applying all cut scores simultaneously yields an average consistency of around low 80s to low 90s. The classification accuracy examinations show that the accuracy indexes at each cut score are generally in the upper 90s (see Table 10).

The program also computes the false negative rates for the test, which in effect are an estimate of those students that may have been misclassified in a performance category lower than their true performance category. The results of the false negatives, found in Table 11, indicate that a very small number of students may have been negatively misclassified in this way. In contrast, the false positive rates, which are estimate of those students that may have been misclassified to a performance category higher than their true performance category, are presented in Table 12. The results indicate that a very small number of students may have been positively misclassified.

Table 11: False Negative Classification

Content Area	Overall False Negative	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.077	0.020	0.014	0.017	0.027
Reading	0.060	0.023	0.025	0.023	0.009
Science	0.079	0.013	0.016	0.024	0.027
Writing	0.060	0.040	0.010	0.022	0.011

Table 12: False Positive Classification

Content Area	Overall False Positive	Cut Below Basic - Basic	Cut Basic - Proficient	Cut Proficient - Goal	Cut Goal - Advanced
Mathematics	0.082	0.023	0.024	0.022	0.013
Reading	0.085	0.052	0.020	0.005	0.014
Science	0.077	0.016	0.016	0.020	0.025
Writing	0.025	0.004	0.011	0.005	0.012

Part 6: CAPT3 Standards

When standards were being established for first generation CAPT, a judgmental standard setting process called Modified Angoff (1971) was employed. Through that process, groups of educators who were familiar with the performance of students at a particular grade level in a particular content area were asked to predict how students who just meet a particular standard (e.g., goal standard) would perform on many different CAPT items. Using the judgment of these groups of educators in consideration with other validity checks, appropriate state goal and remedial standards were recommended by the Department and adopted by the State Board of Education. For the second generation CAPT (CAPT2), the standards were set using a method called Book Mark. In the procedure, all items in the test are arranged from easiest to most difficult. Then a group of educators are asked to mark up to the item at which a student at specific standard could respond to correctly. As in the first generation, the standards set by using the Book Mark method were adopted by the State Board of Education.

The third generation (CAPT3) standards were developed by carrying over the CAPT2 standards as well as department staff working with a CAPT3 Standards Advisory Panel composed of technical experts, district content experts and district research and testing specialists. The CAPT3 standards were set to be as rigorous as the CAPT2 standards. Transferring the standards allowed the Department to maintain the same performance standards for NCLB purposes. The purpose of this section is to summarize the procedures used to accomplish the task of carrying over the standards (see Cizek and Bunch, 2007, for a discussion of standard setting procedures).

In all content areas, the standards define the different academic performance levels. The state goal has been an important benchmark for judging the quality of education in Connecticut for more than a decade. The proficient standard is used for accountability purposes as required by No Child Left Behind (NCLB) to make determinations about Adequate Yearly Progress (AYP) and schools in need of improvement.

To continue to comply with the NCLB accountability requirements, the Connecticut State Department of Education (CSDE) carried over from the CAPT2 to the CAPT3 the following previously adopted achievement standards: Below Basic, Basic, Proficient, Goal and Advanced. The process of carrying over the standards was accomplished with an intergeneration linking study which included the equating of CAPT2 forms and CAPT3 forms. In addition to statistically linking the test generations, historical results from past CAPT2 administrations were taken into consideration as well as input from the CAPT Standards Review Panel composed of a diverse group of Connecticut educators, including curriculum directors, teachers and administrators.

The Standards Review Panel assisted in the identification of acceptable and valid test standards for each content area of CAPT3. The CAPT Standards Review Panel was given an overview of the CAPT3 including the content covered, score weighting, and reporting conventions. Differences between CAPT2 and CAPT3 were also discussed. Copies of the complete CAPT3 test booklets were available for reference. In addition, the procedures for carrying CAPT2 standards over to CAPT3 were presented in detail so that committee members would better understand their role in the process. They reviewed data from several related analyses and discussed implications from both an educational perspective and a technical perspective. They were asked particularly to provide input in the following three areas:

- Review the content of the CAPT, score weighting, and reporting conventions.
- Review results from the inter-generational linking procedure to ensure that standards are reasonable and appropriate across content area; and
- Provide subjective input about the reasonableness and consistency of the standards for all content areas based on their content expertise and historical results from past test administrations.

All procedures were discussed with and approved by the Technical Advisory Committee (TAC) prior to implementation. The TAC is composed of nationally recognized experts in the measurement field. Finally, standards proposed by the standards review panel were presented to the State Board of Education for final approval. Standards were established based on scale scores (100-400) in four content areas: Mathematics, Science, Reading, and Writing.

Table 13 shows the range of scale scores in each performance category.

Table 13: 2010 CAPT Achievement Levels and Scale Score Ranges

Content Area	Scale Score Ranges				
	Below Basic	Basic	Proficient	Goal	Advanced
Mathematics	100 - 190	191 - 220	221 - 259	260 - 289	290 - 400
Science	100 - 189	190 - 214	215 - 264	265 - 294	295 - 400
Reading	100 - 173	174 - 204	205 - 250	251 - 282	283 - 400
Writing	100 - 181	182 - 209	210 - 249	250 - 285	286 - 400

Part 7: Validity

According to the 1999 AERA, APA, NCME *Standards*, "It is helpful to consider the four phases leading from the original statement of purpose(s) to the final product: (a) delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured; (b) development and evaluation of the test specifications; (c) development, field testing, evaluation, and selection of the items and scoring guides and procedures; and (d) the assembly and evaluation of the test for operational use.

In the development and maintenance of CAPT each of these phases is carefully planned and implemented. The following sections detail the important psychometric procedures undertaken to ensure a strong validity argument for the use and interpretation of CAPT (Kane, 2006; Messick, 1989).

7.1. Content Validity Survey

In order for the CAPT to serve its intended purposes, it is critical that users of the test results be confident that those results are meaningful. The test must measure those competencies that are critical to the decisions the test scores are informing.

A content validation study was conducted to examine the content validity of the CAPT for its intended applications. For this study, a survey of the strands proposed for the second generation CAPT was sent to approximately 4,000 Connecticut educators, parents, and other citizens. The purpose of the survey was to determine 1) the importance of the proposed Mathematics, Science, Reading Across the Disciplines, and Writing Across the Disciplines strands and 2) whether the strands are taught prior to the end of the 10th grade. The respondents characterized the strands as important educational outcomes to which students would be instructed prior to testing.

7.2. Scoring Quality Assurance Procedures Undertaken during Development

Much of the following discussion applies to procedures undertaken during field testing and test construction phases of development work. Of course quality control is applied during the operational administration, but not with the aim of selecting or removing items.

In order to ensure the validity of inferences made from the CAPT tests there are quality control procedures in place for the scoring of the test. One such quality assurance component is to check the MC answer keys for MC items several times prior to test administration and one final time during the first run of live results. Items yielding low point biserial correlations are checked a final time for miskeying.

For constructed-response (CR) items, CAPT staff and contractor staff work with Connecticut educators to establish score boundaries in a process known as "range finding". The score point examples and training sets so established are carried forward into operational scoring and elaborated with new samples of student responses. Reader training lasts up to several days, and readers must qualify by matching scores to several sets of prescored student responses. Once scoring begins, validity packets are used to maintain reader accuracy. These are packets of student responses with scores pre-assigned by CAPT staff and Connecticut educators. Readers periodically receive these packets, and their responses are compared to the pre-assigned scores. If a reader assigns too many discrepant scores, that reader is retrained or removed from the project. Other QA procedures include a 100% second read for the writing prompts (IW). There is a 20% second read for short answer and extended response items in mathematics and reading comprehension.

7.3. Item Quality Analysis Undertaken During Development

Another part of assessing the quality and validity of inferences made from an instrument is to assess the quality of the items on the test. This quality is typically assessed by examining the classical item statistics as well as the potential for item bias. Item bias could lead to invalid inferences made for certain subgroups.

Item specifications. CAPT employs *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999) as a primary source of guidance in the construction, field testing, and documentation of the tests. The introduction to the 1999 *Standards* best describes how those *Standards* are and will be used in the development and evaluation of CAPT tests:

> Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist. (*Standards*, p. 4)

Thus, the terms 'target' and 'goal' are used when referring to various psychometric properties of the tests. For example, while it is a goal of test development for each high school test to have a reliability coefficient of .90 or greater, it is not our intention to scrap a test with a reliability coefficient of .89. Instead, the test results would be published, along with the reliability coefficient and associated standard error of measurement.

Item statistics. Because the CAPT tests are used in making individual decisions about students, they must be very reliable, particularly at cut points (the score points that separate adjacent achievement categories). Target reliability coefficients of .90 (or higher) are therefore set for the important cut points of each test.

Other psychometric properties include item difficulty, item discrimination, and differential item functioning. General statistical targets are provided below:

For Multiple-Choice (MC) Items

Percent correct: greater than or equal to .25
Point biserial correlation with total score: greater than or equal to.20
Mantel-Haenszel: No Category C items (see below)

For Constructed-Response (CR) Items

Difficulty: any level as long as all score points are well represented
Correlation with total score: greater than or equal to.20
Generalized Mantel-Haenszel: No chi-square significant at .05 level of alpha

It should be pointed out that the point biserial correlations for MC items and the correlations for CR items refer to total scores of the field test form with the influence of the item in question removed.

Differential item functioning. The Mantel_Haenszel statistic computes an odds ratio for each item that compares item performance for a reference group and a focal group (for whom bias may be an issue). Specifically, the M-H statistic is a ratio of the probability of success on an item for the reference group to the probability of success on the same item for the reference group. When the ratio is greater than one, the probability of success on the item favors the reference group over the focus group. Note that M-H and other methods for identifying statistical bias are flagging mechanisms that do not necessarily mean that the performance difference is due to unfairness in the item. Instead, the standard procedure is for the bias committee review the items to make a final judgmental determination as to whether or not the item is actually biased.

Since its introduction in the field of epidemiology in 1959, Mantel-Haenszel statistics have been employed by many test developers, and several refinements have been added. Educational Testing Service (ETS) uses the Mantel-Haenszel statistic and calculates a D statistic which permits grouping of test items into three categories (Zieky, 1993). The D statistic is a function of the case-control odds estimator of risk generated by SAS's PROC FREQ. The D statistic is calculated as follows:

1. α = case-control estimate of risk (odds ratio)
2. β = natural log of α
3. $D = -2.35*\beta$

Camilli and Shepard (1994, p. 121) describe three categories of items with respect to D:

A	D does not significantly differ from zero using Mantel-Haenszel chi-square, or D's absolute value is less than 1
B	D significantly differs from 0 and D has either (a) an absolute value less than 1.5 or (b) an absolute value not significantly different from 1
C	D's absolute value is significantly greater than or equal to 1.5

Camilli and Shepard note that Category B items are typically investigated for potential bias, while Category C items are typically removed. Others treat Category C items only as candidates for elimination, pending a reprieve from the committee. In other words, Category C items are considered unusable unless specifically declared usable by the committee. It should be noted that an item that allowed a target group to break out of a pattern of trailing behind the reference group on all other items would tend to fall into Category C. The committee would likely want to keep such an item, in spite of its Mantel-Haenszel status.

DIF occurs when an item shows different results by group (e.g., by race, or sex) that cannot be explained by known differences in the overall achievement levels of the two groups. Overall achievement level is typically taken as scores on an operational test, assuming that the operational test is itself free of bias. While committee members are free to examine all field-tested items, they must review all items with a Category C rating. Unless the committee specifically calls for the inclusion of any such item, that item is removed from the pool.

7.4. Equating Design

A different CAPT form is used each year. In order to ensure that appropriate comparisons can be made from one form of the CAPT to another, test forms must be equivalent to each other. Care must be taken when test items are developed, when items are selected to create forms, when tests are administered, and when tests are scored to keep all conditions as similar as possible for one test form to another. Two important characteristic that must be similar across forms are the content that is measured and the difficulty of the test.

Part 4 of this report details the procedures used to equate and scale the CAPT tests. As mentioned above, three independent groups undertake the analyses and cross-check all analyses and results to ensure accuracy. Connecticut expends great effort and resources to maintain an assessment program that employs high quality psychometric standards and quality assurance.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999*). Standards for educational and psychological testing*. Washington, DC: AERA.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508-600), Washington, DC: American Council on Research.

Camilli, G., & Shepard, L. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Cizek, G. J., & Bunch, M . B. (2007). *Standard setting: A guide to sstablishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 18-64). Westport, CT: American Council on Eduation/Praeger.

Linacre, J. M., & Wright, B. D. (1993, 2006). *A user's guide to BIGSTEPS*. Chicago, IL: MESA Press.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education/Macmillan Publishing Company.

Winsteps. (1991-2006©). Linacre, John M.

Zieky, M. (1993). Practical questions in the use of DIF statistics in item development. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 337-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

Appendix A: Item Analysis

Mathematics HS17 Item Analysis

Grid-in Items
PC = Proportion Correct
RPB = Point-Biserial correlation

Open-ended Items
Mean = Mean OE score
Corr = Item-total correlation
0 – 3 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	OE	-0.1342	1.52	0.62
2	OE	0.7082	1.14	0.60
3	OE	0.6540	1.06	0.66
4	OE	0.7322	1.07	0.68
5	GR	-1.2821	0.71	0.45
6	GR	1.0989	0.31	0.22
7	GR	0.2417	0.45	0.54
8	GR	0.2399	0.45	0.55
9	GR	1.0325	0.32	0.52
10	GR	0.8375	0.35	0.56
11	GR	1.0380	0.32	0.50
12	GR	-1.2969	0.71	0.50
13	GR	-0.8934	0.65	0.56
14	GR	-0.4307	0.57	0.37
15	GR	0.4635	0.41	0.54
16	GR	0.2094	0.46	0.60
17	OE	-0.5238	1.69	0.59
18	OE	0.2520	1.30	0.74
19	OE	0.2362	1.33	0.70
20	OE	0.4226	1.23	0.70
21	GR	0.0743	0.48	0.51
22	GR	0.4408	0.42	0.43
23	GR	-0.2400	0.54	0.42
24	GR	-0.2140	0.53	0.49
25	GR	-0.3782	0.56	0.63
26	GR	0.0053	0.49	0.53
27	GR	-0.8362	0.64	0.41
28	GR	-0.3951	0.57	0.60
29	GR	0.8172	0.35	0.60

Item	Type	Rasch	PC/Mean	RPB/Corr
30	GR	0.9394	0.33	0.55
31	GR	-0.7778	0.63	0.56
32	GR	-1.3228	0.71	0.44

Science HS17 Item Analysis

Multiple-choice Items
PC = Proportion Correct
RPB = Point-biserial correlation for keyed answer

Open-ended Items
Mean = Mean OE score
Corr = Item-total correlation
0 – 3 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	OE	-0.0242	1.92	0.62
2	OE	-0.4392	2.17	0.50
3	OE	0.4858	1.60	0.56
4	MC	-0.9130	0.78	0.47
5	MC	-0.9213	0.78	0.36
6	MC	-0.9816	0.79	0.25
7	MC	-0.2726	0.67	0.30
8	MC	1.0231	0.41	0.30
9	MC	0.3087	0.56	0.44
10	MC	0.4823	0.52	0.32
11	MC	0.1353	0.59	0.43
12	MC	1.1884	0.38	0.35
13	MC	0.9808	0.42	0.38
14	MC	0.3816	0.54	0.35
15	MC	0.1785	0.59	0.18
16	MC	-0.0427	0.63	0.52
17	MC	-0.6227	0.73	0.30
18	MC	0.6143	0.50	0.32
19	MC	-0.1854	0.66	0.45
20	MC	-0.8048	0.76	0.38
21	MC	-0.0955	0.64	0.38
22	MC	0.1657	0.59	0.34
23	MC	0.0664	0.61	0.22
24	MC	0.6097	0.50	0.28
25	MC	-0.6104	0.73	0.43
26	MC	-0.7620	0.76	0.44
27	MC	0.2948	0.56	0.44
28	MC	0.3845	0.54	0.30
29	MC	0.6386	0.49	0.30
30	MC	0.1774	0.59	0.29
31	MC	0.8507	0.45	0.31

Item	Type	Rasch	PC/Mean	RPB/Corr
32	OE	0.2653	1.76	0.62
33	OE	0.1765	1.81	0.52
34	MC	-0.9224	0.78	0.26
35	MC	-0.1629	0.65	0.31
36	MC	-0.2322	0.67	0.49
37	MC	-1.4604	0.85	0.37
38	MC	0.8654	0.45	0.40
39	MC	-0.0137	0.62	0.27
40	MC	-0.1777	0.66	0.37
41	MC	-0.4392	0.70	0.35
42	MC	1.1326	0.39	0.24
43	MC	-0.2985	0.68	0.55
44	MC	0.5509	0.51	0.42
45	MC	0.1296	0.60	0.47
46	MC	0.1039	0.60	0.26
47	MC	0.0521	0.61	0.32
48	MC	-0.8780	0.78	0.47
49	MC	0.4835	0.52	0.35
50	MC	0.0604	0.61	0.42
51	MC	-0.0433	0.63	0.50
52	MC	-0.3363	0.68	0.37
53	MC	-0.0313	0.63	0.36
54	MC	-0.4500	0.71	0.46
55	MC	0.4900	0.52	0.26
56	MC	-0.0780	0.64	0.52
57	MC	0.8453	0.45	0.35
58	MC	0.4387	0.53	0.36
59	MC	0.4366	0.53	0.50
60	MC	0.9920	0.42	0.38
61	MC	-0.2420	0.67	0.44
62	MC	-0.8544	0.77	0.55
63	MC	0.1938	0.58	0.46
64	MC	0.0862	0.60	0.58
65	MC	-0.4580	0.71	0.53

Reading for Information HS17 Item Analysis

Multiple-choice Items
PC = Proportion Correct
RPB = Point-biserial correlation for keyed answer

Open-ended Items
Mean = Mean OE score
Corr = Item-total correlation
0 – 2 = Percent of students at each score point

Item	Type	Rasch	PC/Mean	RPB/Corr
1	MC	-1.2203	0.77	0.24
2	MC	-0.8614	0.72	0.23
3	MC	-0.3389	0.63	0.21
4	MC	-1.1462	0.76	0.37
5	OE	-0.2121	1.18	0.50
6	OE	1.2687	0.69	0.53
7	MC	0.6012	0.45	0.33
8	MC	-0.2341	0.61	0.38
9	MC	-0.3024	0.62	0.39
10	MC	-1.0866	0.75	0.25
11	OE	-0.2476	1.14	0.51
12	OE	1.2737	0.77	0.57
13	MC	0.0003	0.56	0.45
14	MC	0.2854	0.51	0.43
15	MC	-0.5358	0.66	0.42
16	MC	-0.4551	0.65	0.28
17	OE	0.8817	0.82	0.58
18	OE	1.4802	0.64	0.56

Editing and Revising HS17 Item Analysis

Multiple-choice Items
PC = Proportion Correct
RPB = Point-biserial correlation for keyed answer

Item	Type	Rasch	PC/Mean	RPB/Corr
1	MC	0.7010	0.63	0.23
2	MC	-0.7236	0.84	0.26
3	MC	-1.3984	0.91	0.36
4	MC	-1.2991	0.90	0.29
5	MC	0.5161	0.66	0.40
6	MC	0.2213	0.71	0.32
7	MC	0.5429	0.65	0.37
8	MC	-0.7595	0.85	0.37
9	MC	0.1278	0.73	0.30
10	MC	0.5028	0.66	0.40
11	MC	-1.0703	0.88	0.28
12	MC	-0.4892	0.81	0.42
13	MC	-0.2671	0.79	0.38
14	MC	-0.6365	0.83	0.39
15	MC	0.7602	0.61	0.23
16	MC	-0.5061	0.82	0.33
17	MC	0.5487	0.65	0.32
18	MC	0.7345	0.62	0.32

Response to Literature and Interdisciplinary Writing HS17 Item Analysis

Extended Response
Mean = Mean EX score
Corr = Item-total correlation
2 – 12 = Percent of students at each point

	Type	Rasch	Mean	Corr	2	3	4	5	6	7	8	9	10	11	12
RL	EX	0.8968	6.70	0.62	0.01	0.01	0.08	0.10	0.25	0.19	0.24	0.08	0.03	0.00	0.00
IW1	EX	1.2241	7.66	0.72	0.02	0.01	0.03	0.05	0.13	0.14	0.31	0.16	0.10	0.03	0.01
IW2	EX	1.2382	7.67	0.72	0.02	0.02	0.04	0.06	0.11	0.14	0.28	0.17	0.11	0.04	0.01

Appendix B: Raw, Theta, and Scale Scores

Raw, Theta, and Scale Scores for Mathematics HS17

Raw Score	Theta	Scale Score
0	-5.2659	110
1	-4.0222	145
2	-3.2740	166
3	-2.8151	179
4	-2.4755	189
5	-2.2018	196
6	-1.9700	203
7	-1.7675	208
8	-1.5867	214
9	-1.4227	218
10	-1.2720	222
11	-1.1323	226
12	-1.0019	230
13	-0.8792	233
14	-0.7634	237
15	-0.6534	240
16	-0.5485	243
17	-0.4481	245
18	-0.3517	248
19	-0.2587	251
20	-0.1687	253
21	-0.0812	256
22	0.0042	258
23	0.0878	260
24	0.1702	263
25	0.2514	265
26	0.3321	267
27	0.4126	270
28	0.4931	272
29	0.5742	274
30	0.6563	276
31	0.7398	279
32	0.8252	281
33	0.9131	284
34	1.0042	286
35	1.0992	289
36	1.1992	292
37	1.3052	295
38	1.4188	298
39	1.5420	301
40	1.6773	305
41	1.8283	309
42	2.0001	314
43	2.2006	320
44	2.4429	326
45	2.7512	335
46	3.1790	347
47	3.8969	367
48	5.1198	400

Raw, Theta, and Scale Scores for Science HS17

Raw Score	Theta	Scale Score
0	-5.5901	100
1	-4.3741	100
2	-3.6638	100
3	-3.2413	100
4	-2.9365	100
5	-2.6962	101
6	-2.4967	111
7	-2.3252	119
8	-2.1743	126
9	-2.0390	133
10	-1.9159	139
11	-1.8028	144
12	-1.6977	150
13	-1.5995	154
14	-1.5070	159
15	-1.4193	163
16	-1.3359	167
17	-1.2561	171
18	-1.1796	175
19	-1.1059	178
20	-1.0346	182
21	-0.9656	185
22	-0.8985	188
23	-0.8331	192
24	-0.7694	195
25	-0.7070	198
26	-0.6459	201
27	-0.5858	204
28	-0.5267	206
29	-0.4686	209
30	-0.4111	212
31	-0.3543	215
32	-0.2981	218
33	-0.2423	220
34	-0.1869	223
35	-0.1319	226
36	-0.0770	228
37	-0.0223	231
38	0.0324	234
39	0.0871	236
40	0.1419	239
41	0.1969	242
42	0.2522	244
43	0.3079	247
44	0.3640	250
45	0.4207	253
46	0.4780	255
47	0.5362	258
48	0.5951	261
49	0.6551	264
50	0.7163	267
51	0.7786	270
52	0.8425	273
53	0.9080	276
54	0.9753	279
55	1.0446	283
56	1.1162	286
57	1.1904	290
58	1.2675	294
59	1.3479	298
60	1.4321	302
61	1.5205	306
62	1.6140	311
63	1.7132	315
64	1.8194	321
65	1.9337	326
66	2.0579	332
67	2.1945	339
68	2.3468	346
69	2.5195	355
70	2.7204	364
71	2.9620	376
72	3.2681	391
73	3.6919	400
74	4.4033	400
75	5.6199	400

Raw, Theta, and Scale Scores for Reading HS17

Raw Score	Theta	Scale Score
0	-5.2497	100
1	-4.1084	100
2	-3.4717	105
3	-3.0960	119
4	-2.8189	129
5	-2.5916	137
6	-2.3930	145
7	-2.2125	151
8	-2.0441	157
9	-1.8840	163
10	-1.7302	169
11	-1.5813	174
12	-1.4361	180
13	-1.2939	185
14	-1.1540	190
15	-1.0158	195
16	-0.8787	200
17	-0.7423	205
18	-0.6062	210
19	-0.4701	215
20	-0.3338	220
21	-0.1970	225
22	-0.0594	230
23	0.0793	235
24	0.2195	240
25	0.3618	246
26	0.5069	251
27	0.6558	256
28	0.8094	262
29	0.9688	268
30	1.1350	274
31	1.3091	280
32	1.4918	287
33	1.6833	294
34	1.8837	301
35	2.0926	309
36	2.3099	317
37	2.5359	325
38	2.7721	334
39	3.0214	343
40	3.2875	353
41	3.5746	363
42	3.8859	375
43	4.2220	387
44	4.5846	400
45	4.9865	400
46	5.4717	400
47	6.1985	400
48	7.3840	400

Raw, Theta, and Scale Scores for Writing HS17

Raw Score	Theta	Scale Score
0	-4.7259	100
1	-3.5137	100
2	-2.8143	100
3	-2.4075	109
4	-2.1217	121
5	-1.9022	129
6	-1.7241	137
7	-1.5739	143
8	-1.4435	148
9	-1.3276	153
10	-1.2225	157
11	-1.1258	161
12	-1.0354	164
13	-0.9499	168
14	-0.8682	171
15	-0.7894	174
16	-0.7127	177
17	-0.6375	180
18	-0.5632	183
19	-0.4894	186
20	-0.4158	189
21	-0.3420	192
22	-0.2677	195
23	-0.1925	198
24	-0.1163	201
25	-0.0388	204
26	0.0403	208
27	0.1211	211
28	0.2039	214
29	0.2891	218
30	0.3767	221
31	0.4673	225
32	0.5612	229
33	0.6588	232
34	0.7608	237
35	0.8676	241
36	0.9799	245
37	1.0984	250
38	1.2237	255
39	1.3564	261
40	1.4966	266
41	1.6446	272
42	1.7998	278
43	1.9618	285
44	2.1296	292
45	2.3026	299
46	2.4804	306
47	2.6632	313
48	2.8517	321
49	3.0472	329
50	3.2512	337
51	3.4656	345
52	3.6923	354
53	3.9337	364
54	4.1931	375
55	4.4763	386
56	4.7943	399
57	5.1696	400
58	5.6546	400
59	6.4186	400
60	7.6667	400